1 Motivation

Consider a very simple framework for analysing the causal effect of a single unit in the treatment unit $i = 0$ and two units in the control group $i = 1, 2$. It is assumed that before the intervention at time period $t = T_0$ the units have a joint distribution of the form[1]

$$\boldsymbol{y} = \begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{before } T_0$$

where $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2)'$ and the positive definite covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \boldsymbol{\sigma}_{12}' \\ \boldsymbol{\sigma}_{12} & \boldsymbol{\Sigma}_2 \end{pmatrix}$$

where σ_0^2 is the variance of Y_0, $\boldsymbol{\Sigma_2}$ is a 2×2 covariance matrix of the vector $(Y_1, Y_2)'$ and $\boldsymbol{\sigma}_{12}$ is a 2×1 vector with elements $cov(Y_0, Y_1)$ and $cov(Y_0, Y_2)$.

We are interested to derive the best unbiased forecast of Y_0 given the controls Y_1 and Y_2 which is obtained as

$$\begin{aligned} \widehat{Y}_0^N &= \mu_0 + w_1(Y_1 - \mu_1) + w_2(Y_2 - \mu_2) \\ &= \mu^* + w_1 Y_1 + w_2 Y_2 \end{aligned}$$

where $\mu^* = \mu_0 + w_1\mu_1 + w_2\mu_2$. This result implies that there is no reason to impose the restrictions $w_1 \geq 0$, $w_2 \geq 0$ (positivity) and $w_1 + w_2 = 1$ (adding-up restriction). Furthermore, the construction of the synthetic control should include a constant term, as otherwise the synthetic control may have a different mean. See also Doudchenko and Imbens (2017) for a careful discussion of these restriction.

[1]For the ease of exposition we suppress the time index t as in this section we neglect any dynamic effects which will be considered in the next section.

For illustration assume that

$$y \sim \mathcal{N}\left(\begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1 & 0.1 & 0.4\\0.1 & 1 & 0.5\\0.4 & 0.5 & 1\end{pmatrix}\right)$$

For this example the optimal weights for the synthetic control result as $w_1 = -0.133$, $w_2 = 0.4667$ and $\mu^* = 1 - w_1 - w_2 = 0.667$. Note that w_1 is negative even if all bivariate correlations between the individuals are positive. One may argue that this solution does not make much sense as it is not clear what is means that Y_1 enters the synthetic control with a negative sign. This demonstrates the trade-off between the optimality in a statistical sense and the economic interpretability of the solution.

What happens if we impose the restrictions that all weights are positive and sum up to unity? In this case the restricted optimum yields the linear combination $\widetilde{Y}_0^N = 0.2Y_1 + 0.8Y_1$. Since all units have the same mean, the restricted solution is unbiased, that is, $\mathbb{E}(\widetilde{Y}_0^N) = \mu_0$. The important difference is in the variance of these estimates. For our example we obtain

$$var(Y_0 - \widehat{Y}_0^N) = 0.827$$
$$var(Y_0 - \widetilde{Y}_0^N) = 1.16$$

It is interesting to note that the variance of the restricted estimate is even larger than the unconditional variance of Y_0. This is possible as $(w_1, w_2) = (0, 0)$ is not included in the restricted parameter space.

It is not difficult to see that if Y_0 is not correlated with Y_1 and Y_2, then the optimal estimate boils down to $\widehat{Y}_0^N = \mu_0$ and, therefore, it does not make sense to involve a synthetic control. In microeconometric studies it is usually assumed that the individuals in the treatment group and the individuals in the control group are uncorrelated. In such cases we do not care about constructing a synthetic control. The crucial feature of synthetic control methods is the correlation between the units in the treatment and control group. In macroeconomic applications the variables in the treatment and control groups are typically correlated and it is therefore important to model the relationship between the variables.

In empirical practice it is often the case that the number of pre-intervention time periods T_0 is small and may even be smaller than k, the number of units in the control group. In this cases some kind of regularization is necessary to obtain a reliable estimate of $\widehat{Y}_0^N$. Doudchenko and Imbens (2017) suggest to invoke the elastic net penalty when estimating the weights. Instead of just shrinking the parameters towards zero we adopt a penalty that is flexible enough to produce more reasonable weighting schemes by using the objective function

$$Q(w, \lambda_1, \lambda_2) = \sum_{t=1}^{T_0} \left(y_{0t} - \mu^* - \sum_{i=1}^{k} w_i y_{it} \right)^2 + \lambda_1 \left(\sum_{i=1}^{k} w_i^2 \right) + \lambda_2 \left(1 - \sum_{i=1}^{k} w_i \right)^2$$

The first part of the penalty weighted by the shrinkage parameter λ_1 is the usual regularization penalty that shrinks the weights towards zero. The second part of the penalty forces the sum of the weights towards unity. Note that if $\lambda_1 \to \infty$ and $\lambda_2 \to \infty$, then the weights converge to equal weights $1/k$ which appears to yield a more reasonable target than shrinking merely towards zero. The estimator can easily be estimated as it has an explicit solution given by

$$\widehat{w}|_{\lambda_1, \lambda_2} = (X'X + \lambda_1 I_k + \lambda_2 \mathbf{1}_k \mathbf{1}_k')^{-1} \left(X' y^0 + \lambda_2 \mathbf{1}_k \right).$$

where X is a $T_0 \times k$ matrix that stacks all observations for $t = 1, \ldots, T_0$ and $i = 1, \ldots, k$ and y^0 is a $T_0 \times 1$ vector stacking the T_0 time series observations of Y_0. In practice the shrinkage parameters can be chosen by cross validation, where our experience suggest that optimising subject to the restriction $\lambda_1 = \lambda_2$ reduces the computing time and already produces reasonable estimates.

2 Dynamic models

When modelling macroeconomic time series it is often assumed that the $(k+1) \times 1$ vector of time series $y_t = (Y_{0t}, \ldots, Y_{kt})'$ can be represented by a vector autoregressive model given by

$$\begin{aligned} y_t &= \alpha + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t \\ y_t &= \mu + A(L)(y_{t-1} - \mu) + u_t \end{aligned}$$

where $\mu = (\mu_0, \mu_1, \ldots, y_k)'$, $\mathbb{E}(u_t u_t') = \Sigma$ is a positive definite covariance matrix.

Let us derive the optimal forecast of Y_{0t} conditional on $\mathcal{I}_t = \{Y_{1t}, \dots, Y_{kt}, y_{t-1}, \dots, y_{t-p}\}$. Let Q be the Cholesky factor of the inverse of the covariance matrix such that $\Sigma^{-1} = Q'Q$, where Q is an *upper* triangular matrix[2] such that

$$\begin{aligned} \widehat{Y}_{0t}^N &= \mathbb{E}(Y_0|\mathcal{I}_t) \\ &= \mu_0 + \sum_{i=1}^{k} w_i(Y_{it} - \mu_i) + \beta(L)'(\widetilde{y}_{t-1} - \mu) \end{aligned} \tag{1}$$

where $w_i = q_{i+1}/q_1$, $q = (q_1, q_2, \dots, q_{k+1})'$ is the first row of the matrix Q', $\beta(L) = \beta_1 L + \cdots + \beta_p L^p$, and $\beta_j = w' A_j / w_1$. The vector $\widetilde{y}_t$ results from replacing the treated series by the non-treated counterfactual $\widetilde{y}_t = (\widehat{Y}_{0t}^N, Y_{1t}, \dots, Y_{kt})$, where $\widehat{Y}_{0t}^N = Y_{0t}$ for $t < T_0$. Accordingly the sequence $\widehat{Y}_{0t}^N$ is obtained from a simple recursion.

An important problem with the optimal solution (1) is that it involves $(p+1)(k+1)$ parameters which may be difficult to estimate reliably in practice. We therefore may replace the fully optimal solution by a distributed lag of the synthetic control, that is,

$$\widetilde{y}_{0t}^N = \mu^* + \sum_{\ell=0}^{q} \gamma_\ell w' y_{t-\ell} \tag{2}$$

where the parameters $\gamma_0, \cdots, \gamma_q$ and w are obtained by minimizing the sum of squared residuals $\sum_{t=q+1}^{T_0}(Y_{0t} - \widetilde{y}_{0t}^N)^2$. Since the parameters enter non-linearly in the objective function, the minimum can be obtained by applying a simple "switching algorithm" where the updates for $\gamma_1, \dots, \gamma_q$ are obtained by running a regression of Y_{0t} on $\widehat{w}' y_t, \dots, \widehat{w}' y_{t-q}$ and a constant. The update of w is obtained by running a regression of Y_{0t} on the vector $\sum_{\ell=0}^{q} \widehat{\gamma}_\ell (Y_{1,t-\ell}, \dots, Y_{k,t-\ell})'$.

[2] Note that Y_0 is the first variable in the vector y_t such that we have to rearrange the usual lower diagonal Cholesky matrix into an upper triangular matrix.